Distributed Multicast Caching Technique

BACKGROUND OF THE INVENTION

1. Field of the Invention.

This invention relates to multicast transmission of information across a data network. More particularly this invention relates to an improved caching technique for use in multicasting over a data network.

2. Description of the Related Art.

The push model for distributing data over the Internet and other client server networks has become more widespread in recent years. In modern versions of this model a server "multicasts" data to an interested subset of clients on the network, known as a "multicast group". Whoever is interested becomes a listener by joining the group.

By their nature, push applications are closer to the broadcasting paradigm of radio and television than to the interactive paradigm of the World Wide Web. As such, broadband networks, such as cable TV or satellite, can be used as a very efficient medium for the transmission of "pushed data". Unfortunately, currently these networks are one-way only. That is to say, data such as a television program is sent from a broadcasting facility (the head-end) to several receivers (end-users) without any feedback. As such, these networks are inappropriate for popular interactive push applications since the latter require a return channel. Although attempts to upgrade

the current public network infrastructure are underway in several places around the world, it will take some years until reliable two-way broadband networks are commonplace and therefore, a mechanism for multicasting over one-way broadband networks is desirable.

Caching systems have been developed to improve the efficiency of data transmission across the internet. Current internet caching systems, however, are based on the unicast TCP/IP transport protocol. Various technical issues have impeded the development of caching systems adapted to multicast transmission. Using the unicast transport protocol, a document is required to be delivered separately to each requesting client of the cache. For example, when two clients request the same documents simultaneously, or within a short interval, the cache transmitter is required to read the document twice, and send it twice. Consequently the resources of both the cache transmitter and the internet are used inefficiently.

The document, *Reliable Multicast Transport Protocol,* Shioshita, Teruji *et al.*, Draft Document for the 37th IETF, Feb. 7, 1997, proposes a transport control mechanism to enable reliable multicast data transfer to a large number of receivers on a TCP/IP network from a server in parallel. This protocol promotes short delivery time, as the data is transferred only once, and conserves bandwidth because only one copy of the data is sent to the server. It has the advantage of requiring only a single session regardless of the number of receivers. How-

ever, despite some optimizations, there remains a requirement for receiver confirmation by ACK/NAK responses and the retransmission of data to selected receivers based on the information associated with the NAK response are disadvantages, as large numbers of receivers issuing ACK/NAK responses can still cause network congestion.

Another known multicast transport protocol is proposed in *Starburst Multicast File Transfer Protocol (MFTP) Specification.* Miller, K. *et al.*, Internet Draft, April 1998. This protocol operates in the Application Layer.

In copending U.S. Application No. 09/138,994, filed Aug. 24, 1998, of common assignee herewith, and hereby incorporated by reference, a technique of IP multicasting over existing broadband networks without using a return link is disclosed. This technique allows the issues of multicast group membership and error detection and recovery to be handled locally within an end-user terminal, without need for returning data to a host. According to the technique a single data transmitter sends a group of data items to a subset of possible receivers over a one-way channel. Each data item is divided into blocks which are encapsulated to form datagrams, each including a block sequence number, a data item identifier, and a timestamp indicating the age of the data item. A group directory is regularly sent by the transmitter to each of the possible receivers. The group directory contains information for all groups of data items, enabling each receiver to select the group of data item it wishes to re-

ceive. Reliability is provided by periodic retransmission of missing data. Despite these advantages, significant problems remain.

SUMMARY OF THE INVENTION

It is a primary advantage of some aspects of the present invention that there is improved caching of content that is multicast across a data network.

A caching arrangement for the content of multicast transmission across a data network utilizes a first cache, which receives content from one or more content providers. Using the REMADE protocol, the first cache constructs a group directory. The first cache forms the root of a multilevel hierarchical tree. In accordance with configuration parameters, the first cache transmits the group directory to a plurality of subsidiary caches. The subsidiary caches may reorganize the group directory, and relay it to a lower level of subsidiary caches. The process is recursive, until a multicast group of end-user clients is reached. Requests for content by the end-user clients are received by the lowest level cache, and should the content not be available, the request is forwarded toward the root of the tree until it is found. The content is then returned to the requestors. Various levels of caches retain the group directory and content according to configuration options, which can be adaptive to changing conditions such as demand, loading, and the like. The behavior of the caches may optionally be modified by the policies of the content providers.

It is an advantage of some aspects of the invention that content need only be transmitted once to multiple receivers.

The invention provides a method of transmitting data over a communications network which includes receiving content from a content provider, and responsive to the content establishing a first group directory in a cache. The method includes transmitting the first group directory from the cache on a data channel to a subsidiary cache, establishing a second group directory in the subsidiary cache, in which the second group directory is derived from the first group directory, and transmitting the second group directory from the subsidiary cache to a multicast group of receivers.

According to an aspect of the invention, the first group directory is transmitted using the REMADE protocol.

According to still another aspect of the invention, the first group directory is transmitted periodically.

According to an additional aspect of the invention, the first group directory is transmitted in response to a request from a receiver.

According to another aspect of the invention, the first group directory is transmitted according to a policy of the content provider.

According to an aspect of the invention, the second group directory is transmitted periodically.

According to still another aspect of the invention, the second group directory is transmitted in response to a request from a receiver.

According to an aspect of the invention, the second group directory is transmitted using a REMADE protocol.

According to yet another aspect of the invention, the second group directory is transmitted according to a policy of the content provider.

According to a further aspect of the invention, the content provider is a plurality of content providers.

According to another aspect of the invention, the subsidiary cache is a plurality of subsidiary caches.

According to a further aspect of the invention, the cache and the subsidiary caches are linked together as a hierarchical tree, the cache forming a root of the hierarchical tree.

Still another aspect of the invention includes receiving a transmission request from a member of the group of receivers, wherein the transmission request is responsive to the second group directory, and responsive to the transmission request, transmitting a data item from the subsidiary cache to the receiver.

According to still another aspect of the invention, the first group directory includes a root directory hierarchically linked to a plurality of subdirectories. The subdirectories carry a list of data items. A subtree of the first group directory is defined by one of the subdirectories and at least one linked subdirectory thereunder.

According to yet another aspect of the invention, the second group directory includes a root directory hierarchically linked to a plurality of subdirectories. The

subdirectories carry a list of data items. A subtree of the second group directory is defined by one of the subdirectories and at least one linked subdirectory thereunder.

The invention provides a computer software product, comprising a computer-readable medium in which computer program instructions are stored, which instructions, when read by at least one computer, causes the computer to execute a method of transmitting data over a communications network. The method includes receiving content in a first server from a content provider, and responsive to the content establishing a first group directory in a cache of the first server. The method further includes transmitting the first group directory from the cache on a data channel to a second server that has a subsidiary cache, establishing a second group directory in the subsidiary cache, wherein the second group directory is derived from the first group directory, and transmitting the second group directory from the subsidiary cache to a multicast group of receivers.

The invention provides a system for transmitting data over a communications network, which includes a first server, having a cache therein, The first server receives content from a content provider, and responsive to the content, establishes a first group directory in its cache, and transmits the first group directory to a second server having a subsidiary cache.

According to an aspect of the invention, the cache and the subsidiary caches are linked together as a hier-

archical tree, the cache forming a root of the hierarchical tree.

BRIEF DESCRIPTION OF THE DRAWING

For a better understanding of these and other objects of the present invention, reference is made to the detailed description of the invention, by way of example, which is to be read in conjunction with the following drawings, wherein:

Fig. 1 is a schematic diagram of a multicasting cache arrangement;

Fig. 2 is a block diagram of a group directory; and

Fig. 3 is a schematic diagram of a multicasting cache arrangement in accordance with the invention that employs a hierarchical tree structure.

DESCRIPTION OF THE PREFERRED EMBODIMENT

In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances well known circuits, control logic, and the details of computer program instructions for conventional algorithms and processes have not been shown in detail in order not to unnecessarily obscure the present invention.

Software programming code, which embodies the present invention, is typically stored in permanent storage of some type, such as a computer readable medium. In a client/server environment, such software programming code

may be stored on the client or a server. The software programming code may be embodied on any of a variety of known media for use with a data processing system, such as a diskette, or hard drive, or CD-ROM. The code may be distributed on such media, or may be distributed to users from the memory or storage of one computer system over a network of some type to other computer systems for use by users of such other systems. The techniques and methods for embodying software program code on physical media and/or distributing software code via networks are well known and will not be further discussed herein.

Turning now to the drawings, and to Fig. 1 thereof, there is shown the high level architecture of a first embodiment of a multicasting system 10 which can be employed using the techniques of the present invention. The source of the files is a content provider 12, which is provided with transmitting capability, but may lack receiving capability. The content provider 12 is a parent with respect to a downstream cache 14. Content is multicast via a data network 16, which may be the Internet, to a plurality of end-user clients 18.

It is possible for the cache 14 to receive content from a plurality of content providers. For example both the content provider 12 and the content provider 20 may submit a catalog to the cache 14, which then combines information from the two catalogs to formulate its own catalog or group directory for subsequent multicast.

The cache 14 employs the REMADE protocol to multicast the content. The REMADE protocol is disclosed in the

above noted U.S. Application No. 09/138,994. The REMADE protocol is a technique of IP multicasting over existing broadband networks without using a return link. This technique allows the issues of multicast group membership and error detection and recovery to be handled locally within an end-user terminal, without need for returning data to a host. According to the technique, a single data transmitter sends a group of data items to a subset of possible receivers. Each data item is divided into blocks, which are encapsulated to form datagrams, each including a block sequence number, a data item identifier, and a timestamp indicating the age of the data item. A catalog, comprising a group directory is regularly sent by the transmitter to each of the possible receivers. The group directory contains information for all groups of data items, enabling each receiver to select the group of data item it wishes to receive. Reliability may be provided by periodic retransmission of missing data.

In some embodiments, improvements in the REMADE protocol which were disclosed in our copending Application No. 09/564,387, filed May 3, 2000, and herein incorporated by reference, may be used in the practice of the present invention.

Referring now to Fig. 2, a representative group directory, directory system 30, is shown. The directory system 30 is a preferably a hierarchical arrangement, and comprises a root directory 32, with links to subdirectories 34. The root directory 32, and any of the subdirec-

tories 34 may have links to general data items 36, or to patches 38. The arrangement is recursive, with subdirectories 34 having links to more deeply nested subdirectories, such as subdirectory 40.

Referring again to Fig. 1, functionally the cache 14 operates according to a service policy that is appropriate for a particular application. Files may be retained or erased, in whole or in part, following a multicast, depending on various factors, which may include file size, the characteristics of the provider or the recipient, average demand for the content, currently available free space, and many other relevant parameters.

As shown in Fig. 3, a second embodiment of the invention employs a multilevel hierarchy 22 of cache servers. As in the first embodiment, a plurality of content providers 12, 20 provide content to a high level cache 24, The content providers 12, 20 have transmitting capability, but may lack receiving capability. In addition to providing content, the content providers 12, 20 optionally act as policy control servers, specifying the recipients to whom the content may be delivered, and the manner of delivery. In default of a content delivery policy received from either of the content providers 12, 20, the cache 24 executes an internal policy. In any case the interaction of the cache 24 and other levels of the multilevel hierarchy 22 is preferably determined by internal configuration options, rather than by the content providers 12, 20.

Upon request, or in accordance with its policy control, the cache 24 delivers catalog and content in accordance with the REMADE protocol over a data network, which may be the Internet, to subsidiary caches 26. The subsidiary caches 26 have both receiving and transmitting capabilities, and depending upon the attributes of their respective clients, may independently reorganize the catalog that was received from the cache 24. In Fig. 3, each of the subsidiary caches 26 has end-user clients 28. It should be noted that while the multilevel hierarchy 22 is represented for clarity in Fig. 3 as a two-level tree, comprising the cache 24 and the subsidiary caches 26, there can be any number of levels in the tree-structured hierarchy, as appropriate for a particular application. In such case, the subsidiary caches 26 communicate with a lower level of caches, rather than directly with the end-user clients 28.

Typically, the end-user clients 28 have both transmitting and receiving capability. When the subsidiary caches 26 have organized their data inventory into a tree-structured catalog according to the REMADE protocol, as disclosed more fully in the above noted Application No. 09/138,994, they transmit it to all the end-user clients 28, or to a predefined multicast group of the end-user clients 28. The end-user clients 28 receive a relevant part of the catalog, or may receive the whole catalog. They choose a document, and begin receiving it according to the REMADE protocol. Of course, various members end-user clients 28 may choose different documents,

in which case all the documents are transmitted according to the governing policy. Clients not currently members of the multicast group may in some circumstances elect to join it. Based on considerations such as the average number of requests for particular content specified in the catalog, the subsidiary caches 26 can independently decide to elect a periodic mode of transmission of the catalog or the content, or to transmit either or both of them on demand. In like manner, the cache 24 can elect a mode of transmission of its catalog to the subsidiary caches 26.

In the multilevel hierarchy 22, should a particular one of the subsidiary caches 26 lack a file, or portion of a file, requested by a receiver, such as one or more of the end-user clients 28, it obtains the missing parts from the cache 24. These parts are immediately resent to the receiver, which considerably reduces latency from the point of view of the receiver.

In a downstream push mode of operation, if a particular content is designated according to the service policy as content in high demand, then the cache 14 may, even in the absence of a request from any of the end-user clients 18, multicast the content. In this mode, the content is flagged, requiring any downstream caches to begin receiving the content immediately, without waiting for transmission requests from clients. In the case of a multilevel cache hierarchy, the caches in each level may push the content down to other levels. As in the other modes of operation, the behavior of the caches at all

levels is controlled by configuration parameters, optionally modified by the policies of higher level caches, or of the content providers 12, 20.

While this invention has been explained with reference to the structure disclosed herein, it is not confined to the details set forth and this application is intended to cover any modifications and changes as may come within the scope of the following claims: